J. R. Statist. Soc. A (2005)
168, *Part* 3, *pp.* 469–472

Editorial: Model selection and efficiency—is 'Which model . . .?' the right question?

Statistics as a science and profession has been transformed over the last few decades by fine-tuning its orientation to serve other scientific fields, encouraged by the revolution in computing technology. As a result, it encompasses a vast variety of activities, provides a wide range of careers and is universally accepted as indispensable to the modern information society. These positive aspects go hand in hand with profound weaknesses. We cannot agree on an authoritative definition of our subject, on a short list of its fundamental principles (those of probability theory are insufficient) or on what amounts to good practice in particular settings, and how to promote it.

Model selection, with the associated uncertainty, is an example of practice that is becoming increasingly problematic as powerful computers and convenient software enable us to explore data in ever greater detail. We can inspect how several alternative models fit the studied data set, and settle on one of them. Such a 'final' model and its maximum likelihood (ML) fit (estimates and standard errors, or information equivalent to them) is the centre-piece of the results section of many a report or manuscript, accompanied by model checking and claims of (approximate) unbiasedness and asymptotic efficiency, after confirming that the requisite regularity conditions have been satisfied.

Despite being regarded as respectable, this approach is flawed because it ignores the consequences of model uncertainty. Since the lucid discussions by Draper (1995) and Chatfield (1995), neither research nor practice has paid much attention to this issue. To Bayesians, the topic might be broached constructively by paraphrasing de Finnetti (1974) ('Every probability is conditional') as 'Every *posterior distribution* is conditional'. We usually study the properties of estimators conditionally on the selected model, ruling out the possibility that the selected model might not be valid. After all, the model selection is a random (data-dependent) process, but the (unknown) 'good' model is fixed, being a property of the studied phenomenon and oblivious to our study design and data collection process.

Model selection forces us to make a decision without considering the consequences of the errors that may have been made in the process. We end up putting all our inferential eggs in one unevenly woven basket. Depending on the purpose, an error of one kind may be innocuous or disastrous relative to an error of another kind. The conditional probabilities of these two kinds of error, controlled in hypothesis testing, are often a poor indication of their gravity.

By way of an example, consider the text-book balanced one-way analysis-of-variance (ANOVA) setting (normality and equal variances within groups) with $K = 8$ observations in each of $J = 10$ groups, and two problems:

(a) what is the mean of group A?;
(b) what is the within-group variance?

For either problem, the text-book method sets up the ANOVA table, 'concludes' that the group means are identical or not and proceeds to estimation based on the corresponding model. Choosing the same model for the two problems in this way may be irrational. For problem (a), the choice is between using $K = 8$ or $JK = 80$ observations, incurring some bias in the latter

 0964–1998/05/168469

case. For problem (b), it is between using $JK - J = 70$ or $JK - 1 = 79$ degrees of freedom, also incurring some bias in the latter case. Compared with the tenfold increase in the effective sample size in problem (a), the nine additional degrees of freedom (13%) that could be gained in problem (b) do not deserve as serious a consideration.

By regarding the 'best' model as the missing information, the EM algorithm (Dempster *et al.*, 1977) or data augmentation (Wei and Tanner, 1990) provides a clear diagnosis of the problem in this two-stage procedure (1, select a model; 2, estimate with the selected model). Carrying out each stage as well as possible does not lead to the compound being carried out as well as it could be. Further, the conventionally reported standard errors are akin to their counterparts in the M-step, whereas correctly estimated standard errors take into account the missing information (Louis, 1982).

The strait-jacket of basing all inferences on a single selected model, promoted by both recent and outdated elementary statistics text-books, is compounded by the confusion of the conditional and unconditional sampling distributions. In the ANOVA example, the selected-model-based estimator is neither unbiased nor efficient, and is distinctly non-normally distributed. There is no need for any theory to prove it—for most readers, simulating the estimator with the details of their choice, and using their preferred software and model selection procedure, is not a demanding exercise. Longford (2003) discussed this problem in the wider context of ordinary regression.

Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the two contemplated estimators for problem (a), and s_0^2 and s_1^2 their respective sampling variances. Further, denote by I the outcome of ANOVA; $I = 1$ if $\hat{\theta}_1$ is used and $I = 0$ otherwise. The estimator that is obtained following the ANOVA-based model choice is neither $\hat{\theta}_1$ nor $\hat{\theta}_0$, but $\hat{\theta} = (1 - I)\hat{\theta}_0 + I\hat{\theta}_1$. This is a *mixture* of the two estimators, and its sampling variance is neither s_0^2 nor s_1^2, and not even the corresponding sampling variance $(1 - I)s_0^2 + Is_1^2$.

In more complex modelling, the formal and informal model selection is usually more extensive, accompanied sometimes by (unreported) improvisation. In this setting, the properties of the estimators conditional on the selected model are likely to be very distant from their unconditional (mixture) versions. Using a more efficient test, test of different size, Akaike information criterion, Bayes information criterion or another information criterion does not alleviate the problem, because each of them selects a model without being informed about the purpose to which the model will be put.

The focused information criterion (Claeskens and Hjort, 2003) is an obvious improvement because it selects the model for which the mean-squared error of the estimator of a specified parameter (or, generally, a target) is the smallest. However, model average estimators (Hjort and Claeskens, 2003) have a much greater potential, because they search in a wider space of estimators. In common with Longford (2003) and Longford (2005), chapter 11, they consider all the convex combinations $b_1\hat{\theta}_1 + \ldots + b_M\hat{\theta}_M$ of the M candidate (single-model-based) estimators. The open question is how to set the coefficients b_m; in all except trivial cases the optimal choices must be estimated. The key advantage of this approach is its greater ambition. Model selection aims to match the most efficient of the candidate estimators (and is not good at it). In contrast, the combination of estimators aims to outperform every one of them. Such estimators have proven their worth in empirical Bayes and small area estimation. Instead of choosing one, they harness the strengths of the alternative estimators (unbiasedness and small variance).

The inefficiency of the selected-model-based estimator was exposed by Freeman (1989) in the analysis of crossover clinical trials, in which the treatment effect might be estimated after choosing between models with and without the carry-over effect. Subsequently, Grieve and Senn (1996) suggested that the treatment effect should be estimated *assuming* that the carry-over is

absent. Longford (2001) proposed an alternative in which the carry-over is assumed to be small. It leads to a combination of the estimators: one that assumes the carry-over to be present and the other to be absent.

A closely related instance of problematic practice is the often stated or implied claim of efficiency of ML estimators for finite samples, dropping the qualifier 'asymptotic'. Suppose that there is no model uncertainty and the ML estimator $\hat{\theta}$ is asymptotically efficient for a parameter of interest θ. For a finite sample, the ML estimator $\hat{\theta}^{\dagger}$ based on a submodel is inconsistent and therefore not asymptotically efficient. But the mean-squared error of $\hat{\theta}^{\dagger}$ may be much smaller than $\mathrm{var}(\hat{\theta})$—the substantial variance reduction by using the invalid submodel may be preferred to a modest bias. In brief, we pay a price of statistical inefficiency by insisting on model validity or correctness. And, when model selection is error prone, we do not attain model validity either.

We should bear in mind that models are our (statistical) invention, meant to assist us in the business of making inferences, and so it is rather disingenuous to shelter our inability to deliver work of high quality under the *caveat* 'Of course, if we do not identify the correct model ...'. (Your foot is at fault when a shoe does not fit well.) My argument is not against models—they are indispensable—but against their unprincipled use.

Bayes factors (Kass and Raftery, 1995) avoid model selection by assigning weights to the estimators based on the candidate models, but the assignment is based on the posterior probability of appropriateness of the model. These probabilities are not informed by the subsequent use of the model fit. In real life, we weigh the anticipated consequences of the decisions that we are about to make. That approach is much more rational than limiting the percentage of making the error of one kind in an artificial (null hypothesis) setting or using a measure of evidence for each model as the weight.

Setting this criticism aside, *p*-values are also evaluated conditionally, both on the selected model *and* that we look at no other output of data processing. In practice, we inspect a whole raft of estimates and *p*-values for other hypotheses, and then report and discuss our pick, ignoring the inspection process that we have applied. There is formidable research on testing multiple hypotheses, but we should apply collective introspection: some of us specialize in this problem and most of the rest ignore it outright. Many years ago, John Nelder called the established practice, which is still the norm today, as resulting in a 'junkyard of false positives'. By ignoring model uncertainty, we are building a 'junkyard of unsubstantiated confidence'.

References

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion). *J. R. Statist. Soc.* A, **158**, 419–466.

Claeskens, G. and Hjort, N. L. (2003) The focused information criterion. *J. Am. Statist. Ass.*, **98**, 900–916.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

de Finetti, B. (1974) *Theory of Probability: a Critical Introductory Treatment*, vol. 1 (translated by A. Machí and A. Smith). Chichester: Wiley.

Freeman, P. R. (1989) The performance of the two-stage analysis of two-treatment two-period crossover trial. *Statist. Med.*, **8**, 1421–1432.

Grieve, A. P. and Senn, S. J. (1996) Estimating treatment effects in clinical crossover trials. *J. Biopharm. Statist.*, **8**, 191–233.

Hjort, N. L. and Claeskens, G. (2003) Frequentist model average estimators. *J. Am. Statist. Ass.*, **98**, 879–899.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.

Longford, N. T. (2001) Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statist. Med.*, **20**, 3189–3203.

Longford, N. T. (2003) An alternative to model selection in ordinary regression. *Statist. Comput.*, **13**, 67–80.

Longford, N. T. (2005) *Missing Data and Small-area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Springer. To be published.
Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.
Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.*, **85**, 699–704.

N. T. Longford
SNTL
Leicester